5.

Q 1. Recall the Gauss-Southwell rule for basic descent methods that we saw in class: $\mathbf{d}_k = -\nabla_{i_k} f(\mathbf{x}_k)\mathbf{e}_{i_k}$, where $i_k = \text{argmax}_{1\le i\le n} |\nabla_i f(\mathbf{x}_k)|$ and $\mathbf{e}_{i_k}$ is the vector that has 0 in all coordinates except for i_k, where it equals 1 (it is the i_kth standard basis vector). Same as in the class, we assume that f is L-smooth. Prove that there exists $\alpha > 0$ such that the Gauss-Southwell rule applied for an appropriate step size α_k satisfies:

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

How would you choose α_k? What can you say about the convergence of this method (discuss all three cases we have covered in class: nonconvex and bounded below, convex, strongly convex)? [10pts]

$d_k = -\nabla_{i_k} f(x_k)\, e_{i_k}$ $\qquad x_{k+1} = x_k - \alpha_k [\nabla f(x^k)]_{i_k} e_{i_k}$

Since f is L-smooth:

$$f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|_2^2$$

$$f(x_{k+1}) \le f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f(x_k) e_{i_k}\rangle + \frac{\alpha_k^2 L}{2}[\nabla f(x_k)]_{i_k}^2$$

$$= f(x_k) - \alpha_k [\nabla f(x_k)]_{i_k}^2 + \alpha_k^2 \cdot \frac{L}{2}[\nabla f(x_k)]_{i_k}^2$$

$$= f(x_k) + [\nabla f(x_k)]_{i_k}^2 \left(\frac{L}{2}\alpha_k^2 - \alpha_k\right)$$

$$\le f(x_k) + \frac{1}{n}\|\nabla f(x_k)\|_2^2 \left(-\alpha_k + \frac{L}{2}\alpha_k^2\right)$$

5.

Since we want $f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2$

we have: $\frac{1}{n}\left(-\alpha_k + \frac{L}{2}\alpha_k^2\right) \le -\frac{\alpha_k}{2}$? (min this.)

$$\frac{L}{2}\alpha_k^2 + \left(\frac{n}{2} - 1\right)\alpha_k \le 0$$

$$\alpha_k \in \left(0, \frac{n-2}{L}\right) \quad X$$

0.

Then $f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2$

According to the conclusion from the class.

nonconvex: $\min_{0\le i\le k} \|\nabla f(x_k)\| \le \sqrt{\frac{2(f(x_0) - \bar{f})}{\alpha(k+1)}}$

0.

convex: $f(x_{k+1}) - f(x^*) \le \frac{2\|x^* - x_0\|^2}{\alpha(k+2)}$

0.

Strongly convex: $f(x_{k+1}) - f(x^*) \le (1-m\alpha)^{k+1}\frac{\|x^* - x_0\|^2}{2\alpha}$

0.

14.

Q 2. Exercise 8 from Chapter 3 in Recht-Wright. Notation from the exercise: $\mathbf{x}_\star = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$. You don't need to worry about getting the same constants as stated there, being off by constant factors (up to 4) is fine. [15pts]

8. **Weakly convex optimization.** Let f be a convex function with L-Lipschitz gradients. Assume that we know the true optimal solution lies in a ball of radius R about zero. In this exercise, we will show that minimizing a nearby strongly convex function will quickly produce a solution that is an approximate minimizer of f. Consider running the gradient method on the function

$$f_\epsilon(x) = f(x) + \frac{\epsilon}{2R^2}\|x\|^2$$

initialized at some x_0 with $\|x_0\| \le R$.

(a) Let $x_\star^{(\epsilon)}$ denote an optimal solution of f_ϵ. Is $x_\star^{(\epsilon)}$ unique?

(b) Prove that $f(z) - f(x_\star) \le f_\epsilon(z) - f_\epsilon(x_\star^{(\epsilon)}) + \frac{\epsilon}{2}$.

(c) Prove that for an appropriately chosen stepsize, the gradient method applied to f_ϵ will find a solution such that

$$f_\epsilon(z) - f_\epsilon(x_\star^{(\epsilon)}) \le \frac{\epsilon}{2}$$

in at most

$$\frac{R^2 L}{\epsilon}\log\left(\frac{8R^2L}{\epsilon}\right)$$

iterations. Find a constant stepsize that yields such a convergence rate.

(a) Claim: $f_{\epsilon}(x)$ is a strongly convex function.

$\forall x, y$ $\quad f_\epsilon(y) - f_\epsilon(x) = f(y) - f(x) + \frac{\epsilon}{2R^2}\|y\|_2^2 - \frac{\epsilon}{2R}\|x\|_2^2$

Since f is convex, according to Taylor's theorem,

$f_\epsilon(y) - f_\epsilon(x) \ge \langle \nabla f(x), y-x\rangle + \langle \frac{\epsilon}{R^2}x, y-x\rangle + \frac{\epsilon}{2R^2}\|y-x\|_2^2$

$= \langle \nabla f_\epsilon(x), y-x\rangle + \frac{\epsilon}{2R^2}\|y-x\|_2^2$

we have: $f_\epsilon(y) - f_\epsilon(x) \ge \langle \nabla f_\epsilon(x), y-x\rangle + \frac{m}{2}\|y-x\|_2^2 \quad \forall x, y. \quad m = \frac{\epsilon}{R^2}$

Thus: $f_\epsilon(x)$ is strongly convex.

then $x_\star^{(\epsilon)}$ is an optimal solution of f_ϵ, $x_\star^{(\epsilon)}$ is unique. 5.

(b) Since $f(z) - f_\epsilon(z) = -\frac{\epsilon}{2R^2}\|z\|_2^2 \le 0$ ①

$f_\epsilon(x_\star^{(\epsilon)}) - f(x_\star) \le f_\epsilon(x_\star) - f(x_\star) = \frac{\epsilon}{2R^2}\|x_\star\|_2^2 \le \frac{\epsilon}{2}$ ②

then: ① + ②: $f(z) - f_\epsilon(z) + f_\epsilon(x_\star^{(\epsilon)}) - f(x_\star) \le \frac{\epsilon}{2}$

$f(z) - f(x^*) \le f_\epsilon(z) - f_\epsilon(x_\star^{(\epsilon)}) + \frac{\epsilon}{2}$ 5.

(c) According to Strong convexity of $f(x)$

$$x_{k+1} = x_k - 2\nabla f(x_k)$$

According to the result in the class:

$$f(x_k) - f(x^*) \le (1-m2)^k \frac{\|x^* - x_0\|_2^2}{2\lambda}$$

$$\le (1-m2)^k \frac{2R^2}{2}$$

WANT: $f(x_k) - f(x^*) \le \frac{\varepsilon}{2}$

$$k \ge \frac{1}{2m} \log\left(\frac{4R^2}{2\varepsilon/2}\right)$$

Since $m = \frac{\varepsilon}{R^2}$ $\quad \|x_0 - x^*\|_2^2 \le 4R^2$

$$k \ge \frac{R^2}{2\varepsilon} \log\left(\frac{8R^2}{2\varepsilon}\right)$$

f_ε is no longer L-smooth!

take $\lambda = \frac{1}{L}$

we have $k \ge \frac{R^2 L}{\varepsilon} \log\left(\frac{8R^2 L}{\varepsilon}\right)$

Choose $\lambda = \frac{1}{L}$ as constant step size,

4.

17.

Q 3 (Bregman Divergence)**.** Bregman divergence of a continuously differentiable function $\psi : \mathbb{R}^n \to \mathbb{R}$ is a function of two points defined by

$$D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle .$$

Equivalently, you can view Bregman divergence as the error in the first-order approximation of a function:

$$\psi(\mathbf{x}) = \psi(\mathbf{y}) + \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + D_\psi(\mathbf{x}, \mathbf{y}).$$

(i) What is the Bregman divergence of a simple quadratic function $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$, where $\mathbf{x}_0 \in \mathbb{R}^n$ is a given point? [5pts]

(ii) Given $\mathbf{x}_0 \in \mathbb{R}^n$ and some continuously differentiable $\psi : \mathbb{R}^n - \mathbb{R}$, what is the Bregman divergence of function $\phi(\mathbf{x}) = \psi(\mathbf{x}) + \langle \mathbf{x}_0, \mathbf{x} \rangle$? [5pts]

(iii) Use Part (ii) and the definition of Bregamn divergence to prove the following 3-point identity:

$$(\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n) : \quad D_\psi(\mathbf{x}, \mathbf{y}) = D_\psi(\mathbf{z}, \mathbf{y}) + \langle \nabla\psi(\mathbf{z}) - \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{z} \rangle + D_\psi(\mathbf{x}, \mathbf{z}). \quad \text{[5pts]}$$

(iv) Suppose I give you the following function: $m_k(\mathbf{x}) = \sum_{i=0}^k a_i \psi_i(\mathbf{x})$, where $\psi_i(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_i\|_2^2 + \langle \mathbf{b}_i, \mathbf{x} - \mathbf{x}_i \rangle$, where $\{a_i\}_{i \geq 0}$ is a sequence of positive reals and $\{\mathbf{b}_i\}_{i=0}^k$, $\{\mathbf{x}_i\}_{i=0}^k$ are fixed vectors from $\mathbb{R}^n$. Define $\mathbf{v}_k = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} m_k(\mathbf{x})$ and $A_k = \sum_{i=0}^k a_i$. Using what you have proved so far, prove the following inequality:

$$(\forall \mathbf{x} \in \mathbb{R}^n) : \quad m_{k+1}(\mathbf{x}) \geq m_k(\mathbf{v}_k) + a_{k+1}\psi_{k+1}(\mathbf{x}) + \frac{A_k}{2}\|\mathbf{x} - \mathbf{v}_k\|_2^2. \quad \text{[5pts]}$$

(i) $D_\psi(x,y) = \frac{1}{2}\|x - x_0\|_2^2 - \frac{1}{2}\|y - x_0\|_2^2 - \langle y - x_0, x - y \rangle$

$= \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|x_0\|_2^2 - \langle x, x_0 \rangle - \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|x_0\|_2^2 + \langle y, x_0 \rangle - \langle y, x \rangle$

$+ \langle x_0, x \rangle + \|y\|_2^2 - \langle x_0, y \rangle$

$= \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x \rangle + \|y\|_2^2 = \frac{1}{2}\|x - y\|_2^2$ 5.

(ii) $\phi(x) = \psi(x) + \langle x_0, x \rangle$

$D_\phi(x,y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle = \psi(x) - \psi(y) + \langle x_0, x - y \rangle - \langle \nabla\psi(y) + x_0, x - y \rangle$

$= \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle = D_\psi(x,y)$ 5.

(iii) $D_\psi(x,y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$

$D_\psi(z,y) + \langle \nabla\psi(z) - \nabla\psi(y), x - z \rangle + D_\psi(x,z)$

$= \psi(z) - \psi(y) - \langle \nabla\psi(y), z - y \rangle + \psi(x) - \psi(z) - \langle \nabla\psi(z), x - z \rangle$

$+ \langle \nabla\psi(z) - \nabla\psi(y), x - z \rangle$

$= \psi(x) - \psi(y) - \langle \nabla\psi(y), z - y \rangle - \langle \nabla\psi(y), x - z \rangle$

$= \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$

then we have $D_\psi(x,y) = D_\psi(z,y) + \langle \nabla\psi(z) - \nabla\psi(y), x - z \rangle + D_\psi(x,z)$ 2.

Use (iii)!

(iv) Since $m_k(x) = m_{k+1}(x) - a_{k+1}\psi_{k+1}(x)$

WANT: $m_k(x) \geq m_k(v_k) + \frac{\beta_k}{2}\|x - v_k\|_2^2$ ①

$\psi_i(x) = \frac{1}{2}\|x - x_i\|_2^2 + \langle b_i, x\rangle - \langle b_i, x_i\rangle$

According to (ii): $D_{\psi_i}(x, y) = \frac{1}{2}\|x - y\|_2^2$

$$D_{m_k}(x, v_k) = m_k(x) - m_k(v_k) - \langle \nabla m_k(v_k), x - v_k\rangle$$
$$= \sum_{i=0}^{k} a_i \{\psi_i(x) - \psi_i(v_k) - \langle \nabla \psi_i(v_k), x - v_k\rangle\}$$
$$= \sum_{i=0}^{k} a_i D_{\psi_i}(x, y) = \frac{\beta_k}{2}\|x - v_k\|_2^2 \quad ②$$

Since $\psi_i(x)$ is convex, $m_k(x) = \sum_{i=0}^{k} a_i \psi_i(x)$ $a_i > 0$ $\forall i$

then $m_k(x)$ is convex.

Since $v_k = \operatorname{argmin}_{x \in R^n} m_k(x)$

then $\nabla m_k(v_k) = 0$

Consider $D_{m_k}(x, v_k) = m_k(x) - m_k(v_k) - \langle \nabla m_k(v_k), x - v_k\rangle$

$= m_k(x) - m_k(v_k)$ ③

According to ② and ③, we have:

$$m_k(x) - m_k(v_k) = \frac{\beta_k}{2}\|x - v_k\|_2^2$$

then: $m_{k+1}(x) \geq m_k(v_k) + a_{k+1}\psi_{k+1}(x) + \frac{\beta_k}{2}\|x - v_k\|_2^2$

Q 4. In class, we have analyzed the following variant of Nesterov's method for L-smooth convex optimization:

$$\mathbf{x}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1}$$
$$\mathbf{v}_k = \mathbf{v}_{k-1} - a_k \nabla f(\mathbf{x}_k)/L$$
$$\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k),$$

where L is the smoothness constant of f, $a_0 = A_0 = 1$, $\frac{a_k{}^2}{A_k} = 1$, $A_k = \sum_{i=0}^{k} a_i$. We take $\mathbf{x}_0 \in \mathbb{R}^n$ to be an arbitrary initial point and $\mathbf{y}_0 = \mathbf{v}_0 = \mathbf{x}_0 - \nabla f(\mathbf{x}_0)/L$.

Prove that we can state Nesterov's method in the following equivalent form:

$$\mathbf{x}_k = \mathbf{y}_{k-1} + \frac{a_k}{A_k}\left(\frac{A_{k-1}}{a_{k-1}} - 1\right)(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}),$$
$$\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k). \tag{1}$$

Hint: It is helpful to first prove that $\mathbf{y}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k$. [10pts]

4, Claim ①: $y_k = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_k}{A_k} v_k$

$y_k = x_k - \frac{1}{L}\nabla f(x_k) = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_k}{A_k} v_{k-1} - \frac{1}{L}\nabla f(x_k)$

Since $v_k - v_{k-1} = -a_k \cdot \frac{1}{L}\nabla f(x_k)$

$\frac{a_k}{A_k}(v_k - v_{k-1}) = -\frac{a_k^2}{A_k}\cdot\frac{1}{L}\nabla f(x_k) = -\frac{1}{L}\nabla f(x_k)$

Thus: $y_k = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_k}{A_k} v_{k-1} + \frac{a_k}{A_k}(v_k - v_{k-1}) = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_k}{A_k} v_k$ ∎

Then, $x_k = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_k}{A_k} v_{k-1}$

$= y_{k-1} + \frac{a_k}{A_k}(v_{k-1} - y_{k-1})$

Claim ②: $v_{k-1} - y_{k-1} = \left(\frac{A_{k-1}}{a_{k-1}} - 1\right)(y_{k-1} - y_{k-2})$

According to claim ①: $y_{k-1} - \frac{A_{k-2}}{A_{k-1}} y_{k-2} = \frac{a_{k-1}}{A_{k-1}} v_{k-1}$

$$\Leftrightarrow \quad \frac{A_{k-1}}{a_{k-1}} y_{k-1} - \frac{A_{k-2}}{a_{k-1}} y_{k-2} = v_{k-1}$$

$$v_{k-1} - y_{k-1} = \left(\frac{A_{k-1}}{a_{k-1}} - 1\right) y_{k-1} - \frac{A_{k-2}}{a_{k-1}} y_{k-2}$$

$$= \left(\frac{A_{k-1}}{a_{k-1}} - 1\right) y_{k-1} - \left(\frac{A_{k-1}}{a_{k-1}} - 1\right) y_{k-2}$$

$$= \left(\frac{A_{k-1}}{a_{k-1}} - 1\right)\left(y_{k-1} - y_{k-2}\right) \quad \blacksquare$$

then:

$$\begin{cases} x_k = y_{k-1} + \frac{a_k}{A_k}\left(v_{k-1} - y_{k-1}\right) \\ \quad = y_{k-1} + \frac{a_k}{A_k}\left(\frac{A_{k-1}}{a_{k-1}} - 1\right)\left(y_{k-1} - y_{k-2}\right) \\ y_k = x_k - \frac{1}{L}\nabla f(x_k) \end{cases}$$

Q5) LSD is incorrect

-5